Traffic Engineering for Video-On-Demand Systems

Bill Wall
Scientific-Atlanta

Abstract

Traffic engineering techniques borrowed from the telephony world are used predict the performance of VOD systems. This paper provides a brief overview of traffic theory and how it can be applied to VOD systems. Erlang's B formula is used to calculate service blockage as function of buy rate, system capacity, and service group size. The implications of these calculations are discussed.

THE PROBLEM

Video-On-Demand (VOD) Systems are now a commercial reality and are operational in a number of cable systems. These systems offer the cable operator the potential for significant revenue increase, but also represent a significant capital investment. The objective, of course, is to maximize the revenue for the minimum investment. In VOD systems, individual video *streams* are created for each active user of the system. The incremental capital cost of adding additional stream capacity is relatively linear, and the cost of VOD systems is often measured in *cost per stream*. If a cable system is under provisioned, that is the peak demand for VOD outstrips the system capacity, then customer dissatisfaction with not being able to get on-demand services on demand may actually cause a loss of revenue. On the other hand, over provisioning by having a higher stream capacity than peak demand costs more than is necessary for delivering the service. Traditional wisdom, based on trial experience, has suggested that VOD system stream capacity should be 10% of the potential VOD customer base. For example, if a cable system has 40,000 digital subscribers capable of receiving VOD, then the VOD system should support 4000 individual video streams. One objective of this paper is to examine that premise and its assumptions. The problem then is to determine what is the optimum capacity needed to support user demand, and what is the optimum way to deploy that capacity.

CONCEPT OF SERVICE GROUPS

In the previous example, 4000 video streams were required for the 40,000 digital subscribers. Assume for the moment that these numbers represent a cable system of 100,000 homes passed, 80% cable penetration, and 50% of those subscribers take digital. Also for the moment assume that the 10% peak VOD usage represents the correct VOD deployment for the system. Typical encoding rates for good quality VOD are about 3.5 Mbps. Using 256 QAM in the plant, with a payload of 38 Mbps, each QAM modulator can carry 10 video streams. The capacity for 4000 streams then requires 400 QAM modulators to support the VOD service. If the cable system is built with 500 homes passed/fiber node, then we could distribute two QAM modulators to each of the 200 fiber nodes, using 12 MHz of plant bandwidth. The same two six MHz channels would be used for VOD applications in all nodes. Alternatively we could feed identical signals to two fiber nodes forming a *service group* of 1000 homes passed that are now logically one group. Here we could feed four QAM modulators to each service group and achieve the same 10% stream capacity, but now using 24 MHz of spectrum. In a similar manner we could combine 4 nodes to form service groups of 2000 homes passed and use eight QAM modulators, occupying 48 MHz of bandwidth, and so forth… Clearly the first case used the least spectrum, but are there advantages of larger service groups? Intuitively, larger groups should have some advantage by "averaging" over a larger population of users during peak usage times. Does this advantage exist and can it be quantified? Similar questions have been dealt with for generations in telephony systems using traffic engineering.

FUNDAMENTALS OF TRAFFIC ENGINEERING

Methods for determining how much capacity is required as a function of expected demand are called *traffic engineering methods.* These methods, a branch of applied probability theory, have been developed over years to accurately predict the performance of telephone and other telecommunications networks.

The demand on a traffic system is called the *offered load* and is the product of the average rate of customer requests (*average arrival rate, r*) and the average time they require service (*average hold time, t*), or offered load, *a*, is given by

$$a = rt$$

The value *a* is dimensionless and expressed in *erlangs*, named after the founder of traffic theory, A. K. Erlang. If the instantaneous demand on the system exceeds the capacity of the systems then the call is *blocked.* Two classes of systems are typically used in the telecommunications world. In the first class, when a call is blocked, it is dropped, and the user must retry at a later time. These systems are called *blocked-calls-cleared* (BCC) systems. In the second class, when a call is blocked, it is put in queue to be serviced at a later time; these are called *blocked-call-delayed* (BCD) systems. Hybrids of the two classes are also popular, where a fixed-length queue is used, but when the queue is filled, calls are dropped.

Telephone traffic cannot be predicted exactly, but may be viewed as statistical processes. A common assumption is that the probability of a call arrival during an interval T is proportional only to the length of the interval, and the constant of proportionality is the average arrival rate *r*. This assumption leads to the fact that the probability that k calls arrive in an interval T is described by a *Poisson distribution*, and any process following this distribution is a *Poisson process.* This process gives an accurate description of telephone call arrivals. Calls exiting the network are assumed to follow a similar process, in an interval T, each call will terminate with probability T/*t* where *t* is the average hold time. This leads to a *negative exponential distribution* H(T) denoting the probability of a given call lasting for a duration of T.

Assuming a BCC system that has a capacity of supporting c calls, and a random load *a* is offered, then Erlang showed that the probability B of an arriving call being blocked is given by the formula

$$B(c, a) = \frac{a^c/c!}{\sum_{k=0}^{c} a^k/k!}$$

This equation is often referred to as the Erlang Loss Formula, or Erlang B Formula and is central in the planning of telecommunications systems. Similar formula can be derived for BCD systems, as well as a calculation for average delay in the queue. These results can be found in most traffic engineering texts.

APPLICATIONS TO VOD SYSTEMS

In many ways VOD systems are analogous to telephony systems. A single VOD session connects a client to a server, much like a telephone call connects two users. VOD sessions and telephone calls both use a similar connection procedure; in fact the DSM-CC session set up procedure was loosely based on Q.931 call setup procedure. Both are initiated randomly by a user. Both last some finite time and terminate. The most straightforward way to model a VOD system is as a BCC system, where when the system is busy, service is denied and a user must try again later. BCD systems could be implemented, but with the relatively long average hold time of a VOD movie, queuing delay could be unacceptably long. This is perhaps a topic for further investigation. Service request arrival statistics should be similar to that of a telephony system, and a Poisson distribution should accurately model this process. However call hold time in a VOD system is less likely to be as random as a telephone call due to the deterministic nature of the fixed length of a video program. However users will terminate early, invoke pause and rewind functions that will extend the length of the program, and programs will vary in length. Perhaps a Gaussian or Raleigh distribution would more closely match VOD systems hold time statistics than the negative exponential distribution. Here we have very little data from fielded systems. This area is clearly one where more work needs to be done. With these caveats, we press forward and use the Erlang B formula to calculate blocking probabilities in a service group and explore the blocking percentage as a function of service group size and buy rates. One remaining key issue is how to relate buy rates to peak offered loads.

ESTIMATION OF PEAK BUSY HOUR

Traffic engineering theory generally assumes that the processes are stationary, which means the parameters describing the process (average arrival time and average hold time) are constant or vary slowly compared to the actual call rate. Traffic loads do vary with the time of day and day of week, as well as season, but in general in telephony systems these variations are slow enough that traffic theory works well. In order to provide high reliability and a high level of customer satisfaction, telephone systems are engineered based on the busiest hour of the day in the busiest season. It would seem appropriate to engineer VOD systems to similar criteria. To date, there is limited public data detailing buy rates for VOD systems versus time of day and time of week. For the calculations shown as example in this paper, some assumptions must be made.

The buy rate model used in the following analysis makes two simplistic assumptions, first that all VOD buys occur in a six hour time segment each evening, and second that Saturday night buy rates are double the buy rates of other nights. We know the first assumption overstates the buy rates during primetime, but the second assumption most likely understates the popularity of weekend primetime. In some manner this may come closer to actual peak rates on weekends. This model would predict that 15% of all buys occur during a three hour primetime period on Saturday night. Cable operators are used to thinking of Pay-Per-View in terms of buys per month. Using the above model we relate peak busy hour average arrival time to buys per month. Based on a four week month, and six hour per day buy period yields a 168 hour/month buy opportunity. The average buy rate per hour for *b* buys per month per sub would be b/168 and the peak buy rate would be 2b/168 or 0.012b. This value times the number of subs per service group yields the average arrival rate for peak busy hour for that service group.

CALCULATION OF BLOCKAGE RATES

The first case examined looks at the sensitivity to service group size for a fixed percentage rate of VOD deployment. The baseline assumption is a plant design of 500hp per fiber node, 80% subscriber penetration, and 20% digital penetration. This first case looks at a fixed buy rate of four per month. Table 1 shows the relevant parameters and the calculated blockage rates. Average hold time used was two hours, to correspond with average movie length.

Number of Nodes	Service Group Size	Number of Digital Subscribers	Number of QAMs	Number of Streams	Offered Load	Blockage Probability
1	500	80	1	10	7.6	10.4%
2	1000	160	2	20	15.2	4.9%
3	1500	240	3	30	22.8	2.7%
4	2000	320	4	40	30.4	1.7%
6	3000	480	6	60	45.6	0.7%

Table 1 - Probability of Blocking vs. Service Group Size

Note first that the offered load using this model is just slightly below the suggested 10% VOD provisioning number, which implies that small statistical variations above the offered load would block at that deployment level. The actual deployment level in this example was 12.5% of digital subscribers in order to match up with QAM granularity. Note also the strong dependence of blocking probability with service group size. This result validates our earlier thought that larger service groups would provide better "averaging" of the load. With this level of digital deployment and this buy rate, a service group size of 2000 homes passed has a blocking probability of under 2%

.

Buys/ Month	Offered Load	Blocking Probability		
		30 Streams	40 Streams	50 Streams
3	22.8	2.7%	.032%	3E-5%
4	30.4	13.9%	1.7%	.03%
5	38.4	27.5%	9.5%	1.2%
6	45.6	37.5%	19.3%	6.0%
7	53.3	45.8%	29.0%	14.2%
8	60.8	52.1%	36.8%	22.5%

Table 2 – Blocking Probability

The second case examined looks at blocking probability as a function of buy rate and level of VOD capacity. The parameters examined would correspond to digital deployment of either 20% in a 2000 homes passed service group or 40% digital deployment in a 1000 homes passed service group. Table 2 lists the relevant parameters and the blocking probabilities.

Note the sensitivity to buy rates, which suggests that it would be impractical to provision a system where blocking does not occur. These results can give an operator a feel for the issues involved in planning a VOD system, however because of the assumptions made in determining average arrival rate for peak busy hour, a better model is needed. Before engineering a VOD system based on these methods, real data needs to be collected and used to determine peak busy hour, and predicted blocking rates need to be verified against real data.

CONCLUSION

A method has been described that can be used in helping to engineer the deployment of VOD systems. Before it can reliably be used, the method needs to be verified against field data. Once verified, this method can help operators design VOD systems and make the business tradeoffs in terms of capital expended versus the probability that a customer is denied service when he attempts to use the system. This method can also be used to aid engineering the tradeoff between service group size and spectrum used for VOD services. One result shows that it will be likely that deployed systems will have occasional denial of service during peak periods, and marketing techniques will need to be developed to cope with this fact. Systems with queuing of requests when capacity is full need to be explored as well. This preliminary analysis also suggests that the 10% capacity rule may be on the low side, but more data is needed. Finally, traffic engineering has been used successfully in telephone systems for decades, and should provide an important tool for the cable industry.